

# 用朴素贝叶斯分类器构建\*\*文本分类\*\*模型

**任务：** 分类一个邮件是spam（垃圾邮件）还是ham（正常邮件）。

**数据集：** SMSSpamCollection.txt（第一列是标签，标注当前样本是spam还是ham）

**难点：** 将文本数据转化为模型可以处理的数值特征。

因为文本是非结构化的，长度不一，机器学习模型不能直接处理，需要将其转化为数值特征。

# 用朴素贝叶斯分类器构建\*\*文本分类\*\*模型

提示：

**步骤1：** 将原始文本数据中的句子分割成单词列表，同时过滤掉标点符号和特殊字符，确保每条文本都被分词成标准的单词列表。

**步骤2：** 基于分词结果，构建一个词汇表，并将每条文本的单词表示为稀疏的词频向量。可以设置一个最低频率阈值，只保留在多条文本中出现过的常见单词。

**步骤3：** 将标签转化为数值。将文本标签（如 "spam" 和 "ham"）转化为数值表示，为后续模型的分类任务准备目标变量。

**步骤4：** 将所有生成的特征（如词频向量）组合成一个整体的特征向量，作为模型的输入特征。

# 用朴素贝叶斯分类器构建\*\*文本分类\*\*模型

例

1. Win a free iPhone! Claim now by clicking the link.
2. I do not like the link。

Step 1: 文本清洗和分词

["Win", "a", "free", "iPhone", "Claim", "now", "by", "clicking", "the", "link",  
"I", "do", "not", "like", "the", "link"]

Step 2: 生成词汇表

词汇表: ["Win", "a", "free", "iPhone", "Claim", "now", "by", "clicking",  
"the", "link", "I", "do", "not", "like"]

Step 3: 基于词汇表, 统计每个样本中单词的词频

比如对于: I do not like the link

[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]

每个位置分别对应win, a, free, iphone等单词出现的次数